DEPARTMENT OF STATISTICS

The University of Wisconsin
Madison, Wisconsin

TECHNICAL REPORT NO. 183

November 1968

A COMPARATIVE POWER STUDY OF THE BIVARIATE RANK SUM TEST AND T^2

by

G. K. Bhattacharyya, Richard A. Johnson

and

H. R. Neave

This research was partially supported by the Office of Naval Research through contract Nonr 1202(17) Project No. 042-222 and by the Wisconsin Alumni Research Foundation.

A COMPARATIVE POWER STUDY OF THE BIVARIATE RANK SUM TEST AND T^2

by

G. K. Bhattacharyya , Richard A. Johnson

and

H. R. Neave

The small sample performance of the bivariate two-sample Wilcoxon type rank sum test W is studied by Monte Carlo evaluation of power under shift alternatives in the bivariate normal distribution. Two types of alternatives are considered: (a) shift in only one coordinate and (b) equal amounts of shifts in both coordinates. The estimated power of the W test is compared with the power of Hotelling's T^2 test. Interestingly, it is found that the empirical power of W substantially exceeds the power of T^2 for some normal shift alternatives although the latter is the uniformly most powerful invariant test for the problem.

1. INTRODUCTION

The present work is devoted to an empirical study of the performance of a Wilcoxon type rank sum test, for bivariate shift, introduced by Chatterjee and Sen [4]. The main emphasis is on the comparison of its power with that of Hotelling's T^2 test for bivariate normal shift alternatives. Although the test is an extension of the univariate Wilcoxon test, its application involves a complex conditioning aspect which is absent in the univariate case and may have segregated it from the domain of applied statistics. Therefore the test is described in Section 2 and its use illustrated.

Nothing whatever is known about the small sample power of the bivariate Wilcoxon test W (it is the test R of Chatterjee of Sen [4]). The computation of exact power is extremely difficult due to the involvement of the probabilites of bivariate rank configurations. Consequently we employ Monte Carlo simulation to estimate the power for several bivariate normal alternatives and compare the results with the exact power of the T^2 test.

The Pitman asymptotic relative efficiency (ARE) of W with respect to T^2 depends on the direction in which the shift occurs as well as on the correlation. Bounds on the ARE for normal and other bivariate distributions have been studied by Bickel [2] and Bhattacharyya and Johnson [1]. This large sample measure essentially reflects the relative performance of the tests under local alternatives and is not always a satisfactory guide to the comparative power in small or moderate samples. The ARE values corresponding to the alternatives considered for empirical power are presented in Section 3 for the purpose of showing the manner in which the ARE reflects power.

It is found that over most of the alternatives considered, the empirical power of W does not lag appreciably behind the power of T^2 test which is the uniformly most powerful invariant test for the normal family. What is more interesting, the power of W seems to exceed that of T^2 for certain shift alternatives and certain correlations in the normal distribution. It is an unusual situation where a nonparametric test seems to perform better than the best known parametric test and certainly contrasts with the univariate case where the power of the Wilcoxon test always falls short of the t-test for normal alternatives since the latter is UMP unbiased (for numerical values see [5]). This points to the need for further theoretical studies on the test W.

2. THE BIVARIATE RANK SUM TEST W.

Let (X_i, Y_i), $i = 1, 2, \ldots, m$ and (X_j, Y_j), $j = m+1, \ldots, m+n$ be independent random samples from two bivariate populations with continuous cdf's $F(x,y)$ and $G(x,y)$ respectively. Consider testing the null hypothesis $H_o: F \equiv G$ against the shift alternataves $G(x,y) = F(x-\theta_1, y-\theta_2)$ where $(\theta_1, \theta_2) \neq (0,0)$.

The computation of the test statistic W consists in combining the two samples together and ranking the $N=m+n$ x and y observations separately in increasing order of magnitude. The combined sample rank matrix thus obtained is denoted by

$$\text{(1)} \qquad \underset{\sim}{R} = \left(\begin{array}{cccc:ccc} R_{11} & R_{12} & \cdots & R_{1m} & R_{1m+1} & \cdots & R_{1N} \\ R_{21} & R_{22} & \cdots & R_{2m} & R_{2m+1} & \cdots & R_{2N} \end{array} \right) ,$$

where the dotted line represents a partitioning of the ranks according to the first and the second samples. The test statistic is given by

$$\text{(2)} \qquad W = (1-q^2)^{-1} (W_1^2 - 2q\, W_1 W_2 + W_2^2)$$

where

$$\text{(3)} \qquad \begin{aligned} W_\alpha &= \sum_{i=1}^{m} R_{\alpha i} - m(N+1)/2 , \quad \alpha = 1,2 \\ q &= 12(N^3-N)^{-1} \sum_{i=1}^{N} R_{1i}\,[R_{2i} - (N+1)/2] . \end{aligned}$$

Note that W_1 and W_2 are the Wilcoxon rank sums in terms of the X and Y marginals and q is Spearman's rank correlation calculated from the entire combined sample.

The determination of the rejection region involves partitioning the N columns of $\underset{\sim}{R}$ into two groups of m and n in all $\binom{N}{m}$ possible ways and

recomputing (2) in each case. Let $W'_1 \leq W'_2 \leq \ldots \leq W'_{\binom{N}{m}}$ be the ordered values thus obtained. The level α rejection region of the permutation test consists of the k largest values of W' if $k = \alpha\binom{N}{m}$ is an integer. If k is not an integer, we have to randomize on the boundary or slightly change α to make k an integer. In practice one need not always compute all the $\binom{N}{m}$ values. It is often possible to recognize the column partitions which give the largest $\alpha\binom{N}{m}$ values of W by inspection of $\underset{\sim}{R}$ and only a few trial computations.

W is well defined except for the case $q = \pm 1$. If q happens to be +1, we modify it to $q' = 1-\epsilon$ where ϵ is a very small number (the value used here is $\epsilon = .001$). Similarly if $q = -1$, we replace q in (2) by $q' = -1 + \epsilon$. This in some sense preserves the continuity of W. Unless the bivariate distribution degenerates to a lower dimension, the occurrence of $q = \pm 1$ has very small probability and hence such modification is seldom needed.

Example. In reliability studies, it is often of interest to compare similar systems produced by two competing manufacturers. Suppose a system consists of two dissimilar components arranged in parallel so that the system fails if and only if both the components fail. Even though the arrangement is in parallel, the assumption of complete independence of the functioning of the components is often unrealistic. Thus we assume that the failure times (X, Y) of the two components have the continuous bivariate distribution F(x, y) for one system and G(x, y) for the other and consider testing of the hypothesis $H_0: F \equiv G$ against $G(x,y) = F(x-\theta_1, y-\theta_2)$, $(\theta_1, \theta_2) \neq (0,0)$. Hotelling's T^2 test requires the assumption not only of bivariate normality for F and G but also of equality for their covariance matrices. Such assumptions for life distributions are hard to justify and hence the applicability of the T^2 test is dubious. The W test on the other hand requires no assumption on the forms of F and G or on their covariance matrices. We illustrate its application by considering

the following hypothetical failure data

	System I				System II			
x	13.5	14.0	8.6	13.3	13.2	8.5	7.8	8.3
y	3.5	18.7	4.6	18.3	17.8	16.6	5.7	10.8

The combined sample rank matrix becomes

$$\underset{\sim}{R} = \left(\begin{array}{cccc:cccc} 7 & 8 & 4 & 6 & 5 & 3 & 1 & 2 \\ 1 & 8 & 2 & 7 & 6 & 5 & 3 & 4 \end{array}\right) .$$

Using (2) and (3), we have

$q = 15/42 = .35714$

$W_1 = 7, \quad W_2 = 0$

$V = (1-q^2)W = 49$

The nonrandomized significance level not exceeding .05 is $\alpha = 3/\binom{8}{4} = .04286$ and hence $k = 3$. Since q is constant for all partitions of $\underset{\sim}{R}$, the rejection region can be equivalently given in terms of V. It would consist of the three largest values of V' under all column partitions of $\underset{\sim}{R}$. By inspection and a few trial computations, we see that each of the following sets of four columns of $\underset{\sim}{R}$ yield the largest value of $V'(= 57.1430)$:

[(1)(2)(4)(5)], [(3),(6),(7),(8)], [(2),(4),(5),(6)], [(1),(3),(7),(8)].

The numbers within braces represent the column numbers of $\underset{\sim}{R}$ when they are numbered serially from left to right. The next lower value of V' is the observed value $V = 49$ and therefore H_o is accepted at $\alpha = .04286$ or even at $\alpha = 4/\binom{8}{4} = .05714$.

3. EMPIRICAL POWER OF W AND COMPARISON WITH T^2

In this section, Monte Carlo simulation is employed to estimate the power of the W test for shifts in several bivariate normal distributions and the estimated power is then compared with the exact power of the T^2 test. computed from the tables of Tang [6] and Tiku [7]. Since both the tests are invariant under scale changes in x and y coordinates, without loss of generality we take unit variances in all the normal distributions. Four evenly spaced values 0, .3, .6 and .9 were selected for the correlation parameter ρ.

There are an enormous number of possible choices for the vector shift parameter $\underset{\sim}{\theta} = (\theta_1, \theta_2)$. For simplicity in computation, we consider two types of shift: (a) a shift in only one coordinate, $\theta_1 = 0$ and $\theta_2 = \theta > 0$ and (b) equal amount of shift in each coordinate, $\theta_1 = \theta_2 = \theta^* > 0$. Four different values are chosen for the single shift parameter in each case. Since our objective is to compare W with T^2, we choose the values of θ and θ^* so that there is an even coverage of the range of power of T^2 and also so that the power T^2 can be read directly from the tables [6, 7] without interpolation.

More specifically, it is known that for the alternative $\underset{\sim}{\theta}$ in a bivariate normal distribution with correlation ρ, the statistic $(N-3)\,T^2/2(N-2)$ is distributed as a noncentral F with (2, N-3) degrees of freedom and noncentrality parameter λ where

$$(4) \qquad \lambda^2 = (mn/N)(1-\rho^2)^{-1}(\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2) \; .$$

The power is tabulated in [6, 7] for different N and certain values of $\phi = \lambda 3^{-\frac{1}{2}}$. For each of the combined sample sizes N = 8, 10 and 12, it is found that the ϕ-values 1.0, 1.5, 2.0 and 3.0 cover evenly the range of power of the T^2 test. Thus fixing the values of ϕ, we determine the amounts of shift by

specializing (4) to the cases (a) and (b) mentioned above. That is, θ and θ^* are determined from

$$\begin{aligned} \theta &= \phi[\,3N(1-\rho^2)/mn\,]^{\frac{1}{2}} \\ \theta^* &= \phi[\,3N(1+\rho)/2mn\,]^{\frac{1}{2}} . \end{aligned} \tag{5}$$

A total of 1,000 replications were made for each of the three sample sizes $(m,n) = (4,4)$, $(6,4)$ and $(9,3)$ and for the different alternatives mentioned above. The data were generated by a uniform $(0,1)$ generator combined with a certain transformation. The uniform generator employed was $\{2^{-26} v_i\}$ where

$$v_{i+1} = 5^5 v_i \pmod{2^{26}}$$

with an arbitrary odd starting value v_0 between 0 and 2^{26}. If U_1 and U_2 are independent uniform $(0,1)$ random variables and X and Y are defined by

$$\begin{aligned} X &= (-2\ln U_1)^{\frac{1}{2}} \sin(2\pi U_2) + \rho(1-\rho^2)^{-\frac{1}{2}}(-2\ln U_1)^{\frac{1}{2}}\cos(2\pi U_2) \\ Y &= (-2\ln U_1)^{\frac{1}{2}} \cos(2\pi U_2) , \end{aligned} \tag{6}$$

then (X,Y) is bivariate normal $N(0,0,1,1,\rho)$(c.f. Box and Muller [3]). For a specific (m,n), $N = m+n$ pairs (x,y) were thus generated and the last n of them were shifted according to the two types (a) and (b). The permutation test W was applied with the two nonrandomized significance levels α_1 and α_2 which envelope $\alpha = .05$. The rejection numbers corresponding to α_1 and α_2 were determined from 1,000 replications and the rejection number corresponding to $\alpha = .05$ was then interpolated. The numbers entered in Tables 1-3, when divided by 1,000, give the estimated powers of the W test at $\alpha = .05$. The exact power of the T^2 test corresponding to specified values of ϕ are read

directly from [6,7] and entered in the second row of each table. For the sample size (8,4), linear interpolation in the reciprocal of the degrees of freedom as discussed by Tiku [7] was employed to interpolate power. The rest of each table exhibits the estimated power of the W test for the two types of shift and four values of ρ. Thus each column of a table presents the empirical power of W at different points on a constant power surface of the T^2 test.

As a check on the extent of internal scatter of the estimates of power, the 1,000 replications in each case were run in four groups of 250 each. It was found that the variation of the rejection numbers from group to group was quite small in most cases. As is typical with the simulation study on a permutation test, the most important contributing factor to the computer time was the generation of $\binom{N}{m}$ column partitions of $\underset{\sim}{R}$ and recomputation of the values of W. The time used for the present job was approximately five hours.

Table 1. Exact power of T^2 and the rejection numbers for W in 1,000 replications for the two types of shift.

Sample size: $m = 4$, $n = 4$; significance level $\alpha = .05$

		ϕ			
		1.0	1.5	2.0	3.0
	T^2	.197	.388	.612	.915
$\theta_1=0$, $\theta_2>0$	$\rho=0$	225	458	642	878
	$\rho=.3$	219	414	634	865
	$\rho=.6$	210	375	547	826
	$\rho=.9$	135	215	339	595
$\theta_1=\theta_2>0$	$\rho=0$	242	399	631	886
	$\rho=.3$	266	466	675	893
	$\rho=.6$	282	491	696	894
	$\rho=.9$	296	536	737	921

Table 2. Exact power of T^2 and the rejection numbers for W in 1,000 replications for the two types of shift.

Sample size: m = 6, n = 4, Significance level α=.05

		ϕ			
		1.0	1.5	2.0	3.0
	T^2	.223	.449	.696	.963
$\theta_1=0$, $\theta_2>0$	ρ= 0	218.5	449.5	696.5	948
	ρ=.3	213.5	448.5	699	945
	ρ=.6	185	399	638	920
	ρ=.9	149	261	418	745
$\theta_1=\theta_2>0$	ρ= 0	216.5	417.5	678.5	937.5
	ρ=.3	240	461	703	939
	ρ=.6	253	493	732.5	962.5
	ρ=.9	263.5	520.5	777.5	964.5

Table 3. Exact power of T^2 and the rejection numbers for W in 1,000 replications for the two types of shift.

Sample size: m = 9, n = 3, significance level α = .05

		ϕ			
		1.0	1.5	2.0	3.0
	T^2	.241	.487	.743	.978
$\theta_1=0$, $\theta_2>0$	ρ= 0	221	431	691	950
	ρ=.3	215	431	666	947
	ρ=.6	216	417	643	926
	ρ=.9	189	344	521	838
$\theta_1=\theta_2>0$	ρ= 0	236	433	693	955
	ρ=.3	243	456	701	954
	ρ=.6	245	458	700	963
	ρ=.9	241	475	708	951

Consider also the ARE of W with respect to T^2 which for the bivariate normal family considered depends on the correlation ρ and the two types of shift. For shifts in the direction $\underset{\sim}{\theta} = (\theta_1, \theta_2)$, the ARE is given by (c.f. Chatterjee and Sen [4])

$$\text{(7)} \qquad e_{W:T}(\underset{\sim}{\theta}, \rho) = \frac{3}{\pi} \, \frac{(1-\rho^2)}{(1-\rho_o^2)} \, \frac{(\theta_1^2 - 2\rho_o\theta_1\theta_2 + \theta_2^2)}{(\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2)}$$

where $\rho_o = (6/\pi) \text{ arc sin } (\rho/2)$. For $\rho = 0$, the expression (7) reduces to $3/\pi$ which is the ARE of the univariate Wilcoxon test relative to t-test. The lower and upper bounds of (7) for all $\underset{\sim}{\theta} \neq \underset{\sim}{0}$ and all $-1 < \rho < 1$ are .87 and .97 respectively [2]. With a view to investigating the extent to which the ARE reflects the comparative power in the present situation, we compute its numerical values for the two types of shift and four values of ρ and present these in Table 4.

Table 4. Values of $e_{W:T}(\underset{\sim}{\theta}, \rho)$

ρ	$\underset{\sim}{\theta} = (0,0)$	$\underset{\sim}{\theta} = (0,0)$
0	.9549	.9549
.3	.9473	.9642
.6	.9241	.9658
.9	.8837	.9592

4. CONCLUSIONS

The scope of the present power study is limited in that only three pairs of sample sizes for each of four bivariate normal alternatives are considered. Due to the permutational structure of the W test, an elaboration of the experiment toward higher sample sizes and non-normal bivariate alternatives would demand an exorbitant amount of computer time. However, within the domain of the present study, our general observation is that the power of W does not fall too far behind that of the best parametric test T^2 except for the type (a) shift occurring in the presence of high correlation. In all cases, the power of W increases with increasing ϕ as does the power of T^2.

From Table 1, one observes that the estimated powers of W are somewhat higher than the power of T^2 for the type (a) shift occurring in the presence of small ρ. Also from Tables 1 and 2, it appears that with increasing correlation, the power of W for type (b) shifts tends to appreciably exceed that of the T^2 test. Although one cannot reach too definite conclusions from empirical studies, the strong indication that W has higher power than T^2 in some cases emphasizes the need for more theoretical studies on the W test.

REFERENCES

1. Bhattacharyya, G. K. and Johnson, R. A., (1968). Approach to degeneracy and the efficiency of some multivariate tests. To appear in Ann. Math. Statist. 39.

2. Bickel, P. J., (1965). On some asymptotically nonparametric competitors of Hotelling's T^2. Ann. Math. Statist. 36, 160-173.

3. Box, G. E. P. and Muller, M., (1958). A note on the generation of random normal deviates. Ann. Math. Statist. 29, 610-611.

4. Chatterjee, S. K. and Sen, P. K., (1964). Nonparametric tests for the bivariate two sample location problem. Calcutta Stat. Assoc. Bulletin, 13 No. 49-50, 18-58.

5. Milton, R. C., (1966). Rank order probabilities: Two sample normal shift alternatives. Tech. Report No. 53a, Dept. of Statistics, University of Minnesota.

6. Tang, P. C., (1938). The power function of the analysis of variance tests with tables and illustrations of their use. Stat. Research Memoirs. 2, 126-149.

7. Tiku, M. L., (1967). Tables of the power of the F-test. Journ. of Amer. Statist. Assoc., 62,